Market Basket Analysis Using Association Rules

Mariam Barakat
ma.barakat@nu.edu.eg

Eslam Mohamed
Es.ahmed@nu.edu.eg

Reem Amgad
r.amgad@nu.edu.eg

Samaa Maged
Sa.maged@nu.edu.eg

Abdelrahman Yasser
a.saeed@nu.edu.eg

Yomna Eid
yeid@nu.edu.eg

Walaa Medhat
wmedhat@nu.edu.eg

School of ITCS, Nile University
Sheikh Zayed, Egypt

***Abstract* - Market Basket Analysis (MBA) is a popular data mining technique that reveals patterns and associations among items that are frequently bought together. Therefore, it is used in detecting patterns in customers purchases to be able to make future decisions and recommendations. In this paper, we apply MBA using association rules to a relational dataset from a grocery store chain and conduct a comparative study of three algorithms: Apriori, FP-Growth and Eclat. We compare the performance and accuracy, using support, confidence, and lift. The results showed that FP-Growth had the best performance, as it is faster than Apriori by 3 times, and 168 times faster than Eclat.**

***Keywords*— Market Basket Analysis, Association Rules, Apriori, FP-Growth, Eclat**

I. INTRODUCTION

Data mining is the process of extracting useful information and knowledge from large and complex datasets [1]. One of the most popular techniques in data mining is association rule mining, which discovers hidden relationships among items in a transactional database [2]. Association rules have many applications in various domains, such as market basket analysis, customer segmentation, and fraud detection [2].

Market basket analysis (MBA) is a common application of association rule mining that aims to identify the products or services that are frequently purchased or used together by customers [1]. MBA can help businesses improve their marketing strategies, increase their sales, and enhance their customer satisfaction [1]. However, MBA is not a trivial task, as there are many challenges that need to be addressed, such as the large size of the data, the high number of possible rules, and the quality and usefulness of the rules [3].

One of the main challenges in MBA is to choose an appropriate algorithm for generating association rules from a transactional database. An association rule is an implication of the form X -> Y, where X and Y are sets of items, called itemsets [4]. An itemset is frequent if its occurrence in the database exceeds a user-specified threshold, called minimum support [4]. An association rule is valid if its confidence, which is the ratio of the support of X union Y to the support of X, exceeds another user-specified threshold, called minimum confidence [4]. Additionally, other measures, such as lift and interest, can be used to evaluate the strength and novelty of an association rule [4].

Several algorithms have been proposed for mining association rules from large datasets. The most popular algorithms are Apriori, FP-Growth, and Eclat. These algorithms are based on different principles and techniques for generating frequent itemsets and association rules. Apriori uses a level-wise approach that iteratively generates candidate itemsets and prunes them based on their support [1]. FP-Growth uses a compact data structure called FP-tree to compress the database and extract frequent itemsets without candidate generation [1]. Eclat uses a depth-first search strategy that exploits vertical data representation and set intersection operations to find frequent itemsets [5].

In this paper, we aim to tackle the problem of finding the best association rules for a given dataset by conducting a comparative study between three enhanced versions of Apriori, FP-Growth, and Eclat. We compare the three algorithms using three metrics: support, confidence, and lift.

In this paper, we follow a three-step approach to mine association rules from a large dataset. First, we create a data warehouse for the chosen dataset, which helps us integrate rows easily. Second, we create a transactional table from the ordering table. Third, we run and evaluate the three algorithms for mining

association rules and compare their performance and accuracy.

The rest of the paper is organized as follows. In Section 2, we present a literature review of the existing research on mining association rules and the advancements in the three algorithms. In Section 3, we describe our methodology for conducting the comparative study, including the datasets, and algorithms. In Section 4, we present and discuss our results. In Section 5, we conclude our paper and suggest some future directions.

II. Literature Review

An enhanced version of the Apriori algorithm was proposed in [6], it uses cSupport and rSupport measures to improve the quality of mined patterns. This approach outperforms the conventional Apriori algorithm in terms of runtime and memory use by contrasting it with their algorithm. Another approach was tested by Pandey [7], where the author suggested adopting the frequent pattern tree (FPT) data structure to improve the Apriori algorithm. The authors demonstrated how their method enhances runtime efficiency and minimizes the number of candidate itemsets created when compared to the conventional Apriori technique. Another study [8] enhances the traditional Apriori algorithm for the e-commerce sector by adding fuzzy set theory and considering sales amount. This new approach yields valuable information for decision-makers, surpassing the traditional association rules with new insights.

In addition, an improved version of the FP-Growth algorithm [9] was suggested. It uses a header table configuration to reduce the complexity of the whole frequent pattern tree. The paper illustrates that it outperforms the conventional FP-Growth algorithm in terms of runtime, memory consumption, and the effectiveness of generated rules by contrasting it with their algorithm. Another approach was suggested [10] to enhance the performance and memory consumption of the FP-Growth algorithm using linked lists. The author demonstrated how this approach outperforms the classic FP-Growth algorithm in terms of memory utilization and runtime. The method consists of two main steps. Firstly, scan the database to determine the number of frequent 1-itemsets. Secondly, the identified itemsets are stored in a linked list, in the second scan frequent 2-itemsets are generated, and so forth for higher-itemsets. The research achieves its purpose with only two scans and as a result, this method leads to less memory and processing.

Wan Abu Bakar et al. [11] introduced i-Eclat algorithm, a more performance-enhancing incremental variant of the Eclat method [12], which requires fewer database scans overall. Experimental results demonstrate i-Eclat's superiority, outperforming the conventional Eclat algorithm in terms of runtime performance by nearly 23%. Eclat's exponential time complexity, arising from the intersection of transactions, is addressed using a randomized approach based on the Bloom filter. By leveraging Bloom Filter's capabilities, such as efficient set membership checking and set operations, BloomEclat significantly improves the intersecting process, leading to faster transaction processing and reduced execution time with a slight false positive error. Another effective variation of the Eclat algorithm to enhance its efficiency in finding frequent itemsets has been put forth [13]. The paper proposes a new algorithm called BloomEclat. By utilizing Bloom Filter's efficient set membership checking and set operations, the intersection process gets improved which leads to faster transactions and reduced execution time with a slight false positive error.

In summary, a comparative analysis of the Apriori, FP-Growth, and Eclat algorithms shows that these algorithms have been continuously improved to enhance their performance. Improvements have been made using a variety of methods, including employing new data structures, using less memory, and enhancing runtime efficiency. The choice of algorithm depends on the specific requirements of the application and the characteristics of the dataset [3]. The findings of these studies can assist researchers and practitioners in selecting the appropriate algorithm for their use case.

III. Methodology

The study employs three popular MBA algorithms, Apriori, FP-Growth and Eclat, and compares their performance and results. The primary data source for this study is the Instacart dataset [14], which contains over 3 million online grocery orders from more than 200,000 Instacart users. The dataset is preprocessed to ensure data quality and consistency. The study then applies Apriori, FP-Growth and Eclat to the preprocessed dataset and generates association rules that indicate the likelihood of customers buying certain items together. The study evaluates the time complexity and performance of the association rules using various metrics, such as support, confidence, and lift. The code is available on GitHub [15].

A. Dataset

The dataset consisted of six tables: Aisles, Departments, Products, Orders, Orders_Products_Prior, and Orders_products_Train. These tables contained information about the products, categories, orders, and customers of the online grocery store. However, to apply Association rules mining algorithms such as Apriori, FP-Growth and Eclat, the data needed to be transformed into a transactional format, where each row represented a unique order and

each column contained the products purchased in that order. To achieve this, the Orders_Products_Prior table, seen in Table I, was iterated and grouped by Order_id, and then the product_id values were converted to product names by joining with the Products table seen in Table II. The resulting rows are 131,210 orders. Table III shows an example of the transactional table created from the original dataset. It consists of the Order_id which represents the uniquely identified identifications for each order. And the products this order contained.

TABLE I
ORDERS_PRODUCTS_PRIOR

Order_id	The id of the previous order
product_id	The product id in the previous order
add_to_cart_order	Previous product was purchased in what order
reordered	If the customer has a previous order that contains the product (1 or 0)

TABLE II
PRODUCT TABLE

product_id	Unique id for each product
product_name	Product name for the specific id
aisle_id	The aisle id in which the product exists
department_id	The department id in which the product exists

TABLE III
TRANSACTIONAL TABLE

Order_id	Unique id for orders
Products	A list of products bought by this Order_id.

Furthermore, the study adopted a data warehousing approach to facilitate the creation of the preprocessed transactional table described above. Data warehousing is a process of collecting, organizing, and storing large volumes of data from various sources in a centralized location. The data warehouse enables efficient and flexible analysis and reporting of the data. For this study, PostgreSQL was used as the database management system for the data warehouse. PostgreSQL is an open-source, relational database that supports advanced features such as transactions, concurrency control, and user-defined functions. The data warehouse was designed using a snowflake schema, which is a type of dimensional modeling that normalizes the dimension tables into multiple levels of hierarchy. Figure 1 illustrates the snowflake schema used for this study.

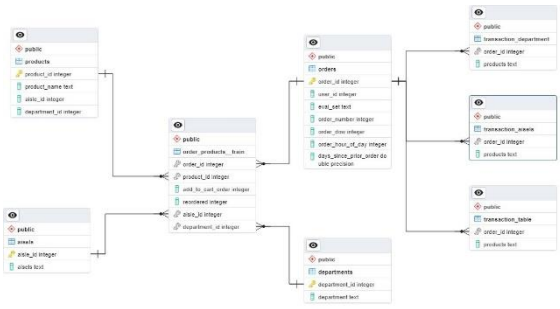

Fig. 1 Snowflake Schema for Instacart data

B. Implementation

Two Python libraries for implementing Apriori, FP-Growth and Eclat algorithms: mlxtend and pyECLAT. Mlxtend was used for Apriori and FP-Growth, it is a library of useful tools and extensions for data science tasks. It provides a function for mining association rules from a transactional database [16]. It also outputs the association rules, support, confidence, and lift of items. These metrics allow us to compare the strength and novelty of the association rules generated by different algorithms. On the other hand, pyECLAT is a simple package for associating variables based on the support of the different items in a dataframe[17]. It is used for the Eclat method. However, this library only outputs the association rules, without the support, confidence, and lift. This makes it hard to compare Eclat with Apriori and FP-Growth.

IV. RESULTS AND DISCUSSION

Apriori, FP-Growth and Eclat were used to analyze the relationships among items that customers often buy together. We found out what products are strongly linked to each other, meaning that buying one product makes customers more likely to buy another product. This information can help businesses improve their product management, placement, and marketing to attract more customers and increase sales.

C. Evaluation Equations

There are important parameters that concern the association rules, namely, support, confidence, and lift [4]. To explain, support is the frequency of an item divided by the number of transactions as shown in eq.1. The X and Y are variables representing different

items, and the N is the total number of transactions. To add, confidence as shown in eq.2 calculates how often Y is purchased when item X is purchased [4]. Lastly, lift is as indicated in eq.3 it showcases the strength of association between two items [14].

$$Support(X \rightarrow Y) = \frac{Freq(X,Y)}{N} = P(X \cup Y) \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{Freq(X,Y)}{Freq(x)} = P(Y \mid X) \quad (2)$$

$$Lift\ (X \rightarrow Y) = \frac{Confid(X \rightarrow Y)}{Freq(x)} = \frac{Confid(X,Y)}{Freq(x)} \quad (3)$$

D. Association Results

Firstly, to find the association rules between products a reasonable confidence was needed to be chosen. For this as shown in figure 2 a graph was plotted between the association rule length versus the confidence value. To be able to compare without having a lot of association rules a confidence threshold of 0.1was chosen.

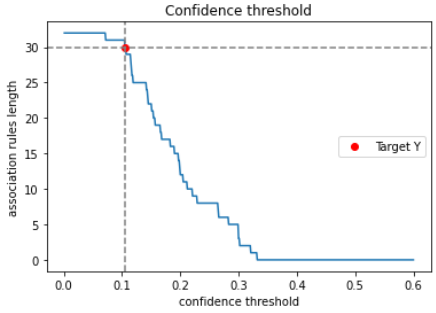

Fig. 2 Choosing 30 products using Confidence threshold.

Figure 3 shows how different products are related to each other based on three measures: support, confidence, and lift. The figure has three clusters of products: blue, green, and red. The blue cluster has low values for all three measures, meaning that the products are weakly related to each other. The green cluster has high support but low confidence and lift, meaning that the products are popular but not strongly related to each other. The red cluster has high confidence and lift but low support, meaning that the products are strongly related to each other but not very common. The most strongly related products are clementines and bags, which belong to the red cluster. Figure 4 displays the same information but in a 3D image. It should be noted that both Apriori and FP-Growth had the same results. Eclat does not provide those metrics.

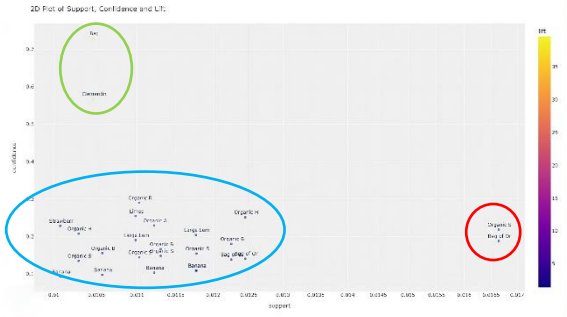

Fig. 3 2D Plot of Support, Confidence, and lift for Apriori and FP-Growth

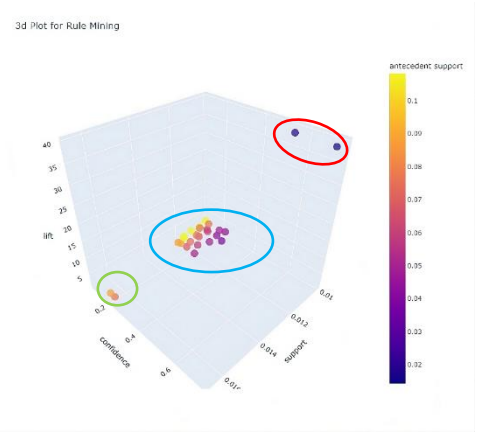

Fig. 4 3D Plot of Support, Confidence, and lift for Apriori

E. Time Complexity

To run these algorithms on a dataset of 131,210 transactions with a minimum support of 0.01. Figure 5 shows how many seconds each algorithm took to run on this dataset. Notice that Eclat took 1,688 seconds, which is 48 times longer than Apriori (34 seconds) and 168 times longer than FP-growth (10 seconds). Therefore, we had reduced the size of the dataset to 500 transactions for Eclat. This result was unexpected, especially for Eclat. One possible reason is that the dataset has a high-density transaction length, which makes the set intersection operation more costly for Eclat algorithm.

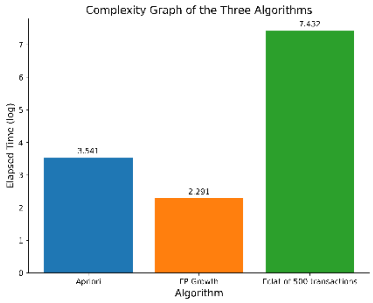

Fig. 5 Time complexity of the three algorithms

V. Conclusion

In conclusion, our analysis of market basket data using different association rule mining algorithms has provided valuable insights into item relationships and purchasing patterns. Among the algorithms we evaluated, FP-growth emerged as the most suitable approach for this dataset. On the other hand, Eclat, despite being a popular algorithm, did not perform as expected in our analysis. The sparse nature of the data may have hindered its ability to extract meaningful associations between items. The lack of support for additional metrics such as lift, and conviction limited our ability to compare Eclat with the other algorithms comprehensively. Overall, FP-growth proved to be a powerful and efficient algorithm for market basket analysis, enabling us to uncover valuable insights into item associations and purchasing patterns. Further research and experimentation could explore ways to improve Eclat's performance on sparse datasets and enhance its suitability for market basket analysis.

References

[1] D. Alcan, K. Ozdemir, B. Ozkan, A. Y. Mucan, and T. Ozcan, "A comparative analysis of apriori and FP-growth algorithms for market basket analysis using multi-level association rule mining," in *Lecture Notes in Management and Industrial Engineering*, Cham: Springer Nature Switzerland, 2023, pp. 128–137.

[2] S. Halim, T. Octavia, and C. Alianto, "Designing facility layout of an amusement arcade using market basket analysis," *Procedia Comput. Sci.*, vol. 161, pp. 623–629, 2019.

[3] L. Hamdad and K. Benatchba, "Association rules mining: Exact, approximate and parallel methods: A survey," *SN Comput. Sci*, vol. 2, no. 6, 2021.

[4] R. Goedegebuure, "Chapter 5 Methods," *Bookdown.org*, 18-Jan-2021. [Online]. Available: https://bookdown.org/robert_statmind/testing3/methods.html. [Accessed: 01-Aug-2023].

[5] J. Korstanje, "The eclat algorithm," *Towards Data Science*, 29-Sep-2021. [Online]. Available: https://towardsdatascience.com/the-eclat-algorithm-8ae3276d2d17. [Accessed: 01-Aug-2023].

[6] S. Tirumalasetty, A. Aruna, A. Padmini, D. V. Sagaru, and A. Tejeswini, "An Enhanced Apriori with Interestingness of Patterns using cSupport and rSupport," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 7, pp. 20–27, 2021.

[7] S. Pandey, "Enhancement of Apriori algorithm for applications of Data Mining using frequent pattern tree," Dublin, National College of Ireland, 2022.

[8] O. Dogan, F. C. Kem, and B. Oztaysi, "Fuzzy association rule mining approach to identify e-commerce product association considering sales amount," *Complex Intell. Syst*, vol. 8, no. 2, pp. 1551–1560, 2022.

[9] M. Shawkat, M. Badawi, S. El-ghamrawy, R. Arnous, and A. El-desoky, "An optimized FP-growth algorithm for discovery of association rules," *J. Supercomput.*, vol. 78, no. 4, pp. 5479–5506, 2022.

[10] S. S. Maw, "An improvement of FP-growth mining algorithm using linked list," in *2020 IEEE Conference on Computer Applications(ICCA)*, 2020.

[11] W. A. Abu, M. Bakar, M. Man, and Z. Man, "I-Eclat: Performance enhancement of Eclat via incremental approach in frequent itemset mining," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 1.

[12] M. J. Zaki, *Fast mining of sequential patterns in very large databases*. New York, 1997.

[13] S. Abbasi and A. Moeini, "BloomEclat: Efficient Eclat Algorithm based on Bloom filter," *Journal of Algorithms and Computation*, vol. 53, no. 1, pp. 197–208, 2021.

[14] "Instacart market basket analysis," *Kaggle.com*. [Online]. Available: http://www.kaggle.com/competitions/instacart-market-basket-analysis/data. [Accessed: 01-Aug-2023].

[15] M. Barakat, E. Ahmed, R. Amgad, S. Maged, and A. Yasser, *Market_basket_analysis: Market basket analysis on Instacart dataset*. .

[16] S. Raschka, "Mlxtend," *Github.io*. [Online]. Available: https://rasbt.github.io/mlxtend/. [Accessed: 01-Aug-2023].

[17] "PyECLAT," *PyPI*. [Online]. Available: https://pypi.org/project/pyECLAT/. [Accessed: 01-Aug-2023].